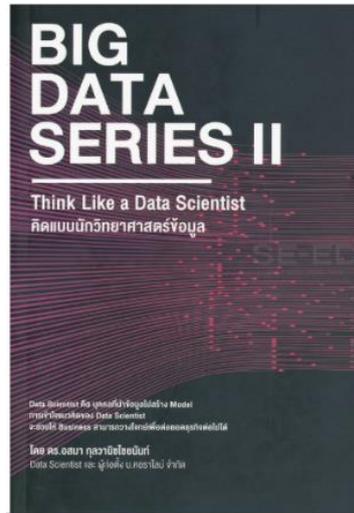


บทวิจารณ์หนังสือ



Big Data Series II

Think Like a Data Scientist

คิดแบบนักวิทยาศาสตร์ข้อมูล

ผู้แต่ง: อสมา กุลวานิชไชยนันท์

ปีที่พิมพ์: 2561 จำนวน 267 หน้า

บทวิจารณ์โดย

ชเนตตี พุ่มพฤษษ์¹

"Big Data Series II: Think Like a Data Scientist" เป็นหนังสือภาคต่อจากหนังสือ "Big Data Series I: Introduction to a Big Data Project" ซึ่งในเล่มแรกได้กล่าวถึง ภาพรวมของกระบวนการจัดทำโครงการ Big Data เอาไว้ สำหรับในเล่มสองนี้ เกิดขึ้นจากการรวบรวมนำเอาประสบการณ์ตรงของผู้เขียน ซึ่งทำงานเป็นนักวิจัยด้าน Big Data ณ ประเทศสหรัฐอเมริกาในช่วงเวลาหนึ่ง และท้ายที่สุดได้ตัดสินใจกลับมาทำงานในอาชีพ Data Scientist ที่ประเทศไทย รวมระยะเวลากว่า 8 ปี หนังสือได้ถ่ายทอดองค์ความรู้จากประสบการณ์จริงด้วยภาษาที่เข้าใจง่าย พร้อมทั้งมีการศึกษาเพื่อให้ผู้อ่านได้เห็นภาพที่ชัดเจน ทั้งแนวทางการใช้งานที่มีการนำเสนอต้นแบบ (Model) ที่เป็นเทคนิคการนำเสนออย่างเหตุผล หนังสือเล่มนี้มีรายละเอียดเกี่ยวกับการวิเคราะห์เชิงลึกและทักษะต่าง ๆ เช่น Data Mining, Machine Learning, Optimization เพื่อหามุมมองในแนวทางที่เรียกว่า Data Scientist จากบทนำของหนังสือเล่มนี้ ผู้เขียนได้แถลงไว้ว่าอาจมีความจำเป็นต้องอ้างอิงคำศัพท์ในเชิงเทคนิคและมีการใช้ภาษาอังกฤษร่วมด้วย แต่หนังสือเล่มนี้ถูกเขียนขึ้นเพื่อคนไทย โดยใช้การวิเคราะห์สถานการณ์ของไทยที่มีวัฒนธรรมการทำงานที่แตกต่าง ผู้เขียนมีการสื่อสารข้อมูลที่

¹ อาจารย์ประจำหลักสูตรบริหารธุรกิจ สาขาวิชาการพัฒนาธุรกิจสร้างสรรค์ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏพระนครศรีอยุธยา, นักศึกษาหลักสูตรบริหารธุรกิจดุซงกีบัณฑิต คณะบริหารธุรกิจ มหาวิทยาลัยรังสิต
e-mail: Chanattee.m@gmail.com

ต่อเนื่องจากเล่มหนึ่ง มากล่าวนำไว้ในบทนำของเล่มสองอย่างน่าสนใจ ทำให้ผู้อ่านเกิดมุมมองทางความคิดเกี่ยวกับ Big Data ง่ายขึ้นและชวนให้ติดตามอ่านตั้งแต่ต้นจนจบ ผู้เขียนได้แบ่งเนื้อหาออกเป็น 7 บทด้วยกัน โดยในแต่ละบทความมีความสำคัญและเชื่อมโยงกันอย่างเป็นระบบนอกจากนี้ผู้เขียนยังมีสรุปท้ายบทไว้ให้ผู้อ่านได้ทบทวนความเข้าใจอย่างชัดเจน โดยเนื้อหาของแต่ละบทมีรายละเอียดสำคัญที่น่าสนใจดังต่อไปนี้

บทที่ 1 Data Ecosystem ผู้เขียนเริ่มต้นด้วยอธิบายถึงระบบนิเวศน์ของข้อมูล (Data ecosystem) ซึ่งเป็นภาพใหญ่ของการทำโครงการ Big Data ซึ่งเป็นบทบาทหน้าที่ของ Data Scientist และยังมีบุคคลที่ไม่ควรมองข้ามอีกหลายส่วน ทั้ง Data Engineer และ Data Analyst ด้วย และยังมีผู้เกี่ยวข้องในงานส่วนอื่น เช่น ผู้จัดการโครงการ (Project Manager) ผู้ใช้งาน ผู้ปฏิบัติงาน (User) เป็นต้น เหตุผลที่ว่าการทำงานต้องประกอบด้วยหลายส่วน เกิดจากงานที่กำลังดำเนินการอยู่นั้น เป็นเรื่องของ Big Data ที่เพียงแค่ชื่อ ก็บ่งบอกถึงความใหญ่อยู่แล้ว จึงนำพาความยุ่งยากในการจัดการ ความสำคัญอยู่ที่การใช้ข้อมูลเหล่านั้นให้เกิดประโยชน์สูงสุด กลไกสำคัญในการทำโครงการ Big Data แต่สิ่งที่เป็นตัววัดความสำเร็จของการทำ Big Data คือ การนำโมเดล (Model) นั้นไปใช้งานได้จริง นอกจากต้องคำนึงถึงข้อมูลที่เก็บมีอะไรบ้าง ต้องใช้ข้อมูลอะไรบ้าง สำคัญก็ตอนนำไปใช้งานนั้นจะใช้งานในรูปแบบไหน ทั้งนี้ การจัดการข้อมูลอย่างง่ายทำให้เห็นภาพชัดเจนที่สุด และเข้าใจได้ง่ายที่สุดว่า ข้อมูลนั้นมีแหล่งกำเนิดได้จากหลายส่วน ทั้งอุปกรณ์ ระบบเก็บข้อมูล ข้อมูลในช่องทางออนไลน์ หรือจะเป็นมนุษย์เราเป็นผู้กรอกข้อมูลเองก็ได้ ซึ่งข้อมูลต่างๆ อาจจะมีความสัมพันธ์กันต้องนำมาพิสูจน์ด้วย Model แต่ก่อนจะนำไปสร้าง Model ได้ ต้องผ่านกระบวนการในการเชื่อมโยงข้อมูลเสียก่อน

บทที่ 2 Waterfall vs Agile วิธีการทำงานแบบเป็นลำดับขั้น โดยใช้หลักการการทำงานแบบ Agile ส่งผลให้เกิดงานที่มีผลลัพธ์ออกมาได้อย่างรวดเร็วภายใต้กรอบงบประมาณ และพร้อมปรับเปลี่ยนไปตามการเปลี่ยนแปลงรอบข้างได้เสมอ ซึ่งการประยุกต์ใช้แนวคิดแบบ Agile สามารถใช้ในกับงานในรูปแบบอื่นด้วยมิได้จำกัดเฉพาะ IT, Big Data หรือการพัฒนาโปรแกรมเท่านั้น ตัวอย่าง เช่น การปรับเปลี่ยนเพื่อสร้างมูลค่าเพิ่มให้ธุรกิจของ Grab อาศัยการ Sprint โดยใช้ Data เข้ามาเป็นหลัก ดังนั้น การทำ Platform เพื่อบริหารอุปสงค์และอุปทานทางธุรกิจ อาจต้องมีการแบ่ง Phase การทำงานเป็นการสร้าง Platform เพื่อเก็บข้อมูลปริมาณการผลิตสินค้าและบริการในแต่ละพื้นที่ ขยายผล Platform ให้สามารถเชื่อมโยงกับอุปกรณ์กับอุปกรณ์ IoT ได้ เพื่อเพิ่มความสามารถในเก็บข้อมูลอย่าง Real-Time เขียนโปรแกรมเพิ่มใน Platform ให้มีการวิเคราะห์ปัจจัยที่มีผลต่อคุณภาพการเพาะปลูก โดยข้อมูลจาก IoT และพัฒนาแนวทางการแสดงผลการวิเคราะห์บน Platform เพื่อให้ User สามารถใช้งานได้อย่างสะดวก และนำ Platform ที่สร้างขึ้นไปใช้งานในวงกว้าง และมีการเขียนนโยบายเพื่อให้เกิดระบบ Digital 4.0 ที่ใช้ได้จริง

บทที่ 3 Data Process Cycle วัฏจักรกระบวนการใช้ข้อมูล เป็นบทที่มีเนื้อหาค่อนข้างมาก เน้นการรวบรวม การทำ Data Process ให้สามารถมองเห็นภาพรวมได้ เพื่อควบคุมระยะเวลาในการดำเนินการได้ตามที่กำหนด ด้วยแรงกดดันของ Big Data จากกระแสสังคม ทำให้เกิดโครงการ Big Data Project หลายโครงการ ที่ต้องอาศัย AI เข้าร่วมด้วย และเก็บข้อมูลด้วย Hadoop นอกจากนั้น ผู้เขียนได้กล่าวถึง มุมมองเชิงธุรกิจที่ใช้ Big Data จาก Deep Learning เพื่อสร้างกลยุทธ์ หรือนำมาทำ Forecasting ต่อไปในวงการเทคโนโลยี และ Artificial Intelligence จะมีการใช้ Deep Learning ค่อนข้างมาก รวมไปถึงวงการการศึกษาและวิชาการก็เช่นกัน พบว่าหัวข้องานวิจัยของนักศึกษาในช่วง 2-3 ปี ที่ผ่านมา มีการเน้นเรื่อง Deep Learning เป็นส่วนใหญ่ ซึ่งถือเป็นเรื่องที่ตีในวงการงานวิจัย รวมไปถึงการนำ Model ไปสร้างระบบ หรือที่เรียกว่า Data Product ทดสอบระบบเสียก่อน เรียกว่า Pilot Test ก่อนใช้งานจริงทั้งระบบ สำหรับ Model ที่

สร้างขึ้นมา สามารถแบ่งประเภทตามการเรียนรู้ 2 ประเภท ได้แก่ Static Model ที่สร้างจากข้อมูลชุดหนึ่ง และเมื่อมีข้อมูลเข้ามาใหม่ จะนำข้อมูลนี้มาใส่ใน Model โดยไม่มีการสร้าง Model ใหม่ ไม่ว่าข้อมูลเข้ามา มากน้อยเพียงใด

บทที่ 4 Optimization Model โมเดลเพื่อหาจุดเหมาะสมภายใต้เงื่อนไขต่างๆ ของ Optimization Model ประกอบด้วย Objective Function เป็นเป้าหมายในรูปแบบการที่ต้องอาศัยการ คำนวณเพิ่มเติม , ตัวแปร (Decision Variable) ที่ต้องการให้โปรแกรมเลือกจุดที่เหมาะสมที่สุด เช่น กรณีการ เลือกเส้นทางการเดินทางด้วย Google Map โปรแกรมจะเลือกเส้นทางที่รวดเร็วภายใต้ข้อจำกัด (Constraints) ของการใช้ Optimization Model ที่มีอยู่รอบตัว ซึ่งมีหลักการทำงาน 4 ขั้นตอน ได้แก่ 1. Schedule Management การวางแผนเพื่อรองรับลูกค้า รวมไปถึงราคาที่เหมาะสม ประเมินจุดคุ้มทุน ปริมาณลูกค้า และความเหมาะสมเรื่องราคา 2. Fleet Assignment ระบุศักยภาพของธุรกิจกับปริมาณความต้องการของลูกค้า และค่าใช้จ่ายต่างๆ ในแต่ละประเภทสินค้าและบริการ 3. Routing การวิเคราะห์ ความสามารถในการให้ได้ประโยชน์สูงสุด 4. Crew Scheduling ต้นทุนพนักงาน การจัดการพนักงานในทุก ขั้นตอน ทั้งยังช่วยในเรื่องข้อมูลปริมาณการสั่งที่เหมาะสมที่สุด (Economic Order Quantity: EOQ) เป็นการ แก้ปัญหาสำหรับการจัดการคลังสินค้า โดยไม่จำเป็นต้องเขียน Model ที่ซับซ้อนแต่อาศัยสมการ EOQ ได้แก่ Ordering Cost ต้นทุนการสั่งซื้อ Holding Cost ต้นทุนการเก็บรักษา และปัจจัยอื่นๆ เช่น Lead Time ในรอ คอยสินค้าของลูกค้า และส่วนท้ายบทได้ให้แนวคิดที่ว่าความคิดนอกกรอบคิดให้แตกต่าง และหาจุดเด่นเป็นการ สร้างคุณค่า

บทที่ 5 Basic Statistics พื้นฐานสถิติ บทนี้เป็นพื้นฐานของบทถัดไป ที่เน้นว่าคนจำเป็นต้องมี ความรู้พื้นฐานด้าน สถิติ ก็คือ การวิเคราะห์ข้อมูล (Data Analyst) และสำคัญที่สุด คือ ต้องบอกได้ว่า วิเคราะห์หอะไร เพื่ออะไร และจะนำผลลัพธ์นั้นไปทำอะไรต่อไป ซึ่งสถิติและทักษะทางคณิตศาสตร์ถือเป็น ประโยชน์ที่สร้างความเติบโตในเชิงข้อมูลเพื่อการตัดสินใจ และนำออกแบบ เป็นกลยุทธ์ทางธุรกิจ สำหรับสถิติ ทั่วไป ได้แก่ ประชากร การหาความสัมพันธ์ การวิเคราะห์การถดถอย หลักสถิติเป็นพื้นฐาน ที่สำคัญอย่างหนึ่ง ของผู้ที่ทำหน้าที่ Data Scientist ความรู้ด้านสถิตินั้นสามารถใช้ได้ในชีวิตประจำวัน

บทที่ 6 Machine Learning การเรียนรู้ของเครื่องจักร ในอนาคตอันใกล้ Machine Learning จะ เป็นเครื่องมือในการทำงานของใครหลายคน โดยมีหลากหลายรูปแบบและยังมีการพัฒนาแนวทางในการสั่ง การ (Algorithm) ให้ Machine ได้ เรียนรู้อย่างต่อเนื่อง ซึ่ง Machine Learning ก็คือโจทย์หรือโครงการ Big Data ที่มีข้อมูลค่อนข้างมาก ก็มีข้อจำกัด ไม่ว่าจะเป็นเรื่องของเวลา เรื่องความเหนื่อยล้า เรื่องประสบการณ์ และมุมมองที่มีต่อข้อมูล ในขณะที่คอมพิวเตอร์สามารถทำงานได้ รวดเร็วและยังรักษาประสิทธิภาพการทำงาน ได้อย่างคงที่ ผู้เขียนเน้นว่า ขั้นตอนพื้นฐานที่สำคัญของการทำ Machine Learning คือการเรียนรู้หลายๆ ครั้ง จนกว่าจะได้ Model ที่เหมาะสม ในการสร้าง Model แต่ละครั้งต้องมีตัวตั้งต้น ที่เรียกว่า Input และจะมี ผลลัพธ์ออกมาเป็น Output ในส่วน Input เป็นหน้าที่หลักของมนุษย์ในการวิเคราะห์ใส่ข้อมูลที่ตี เพื่อให้ได้มา ซึ่ง Model ที่มีคุณภาพ และ Output ที่ออกมามีประสิทธิภาพนั่นเอง

บทที่ 7 Model Evaluation and Implementation การประเมินผลลัพธ์จากโมเดลและการ นำไปใช้งาน ผู้เขียน อธิบายการประเมินผลลัพธ์ไว้หลายรูปแบบ ขึ้นอยู่กับเป้าหมายในการสร้าง Model แบ่ง ออกเป็น 4 ประการ ได้แก่ ความ แม่นยำ ความเร็วในการประมวลผล ความยากลำบากในการใช้งาน ความพึง พอใจในตัว Model ซึ่งอาจมีวิธีการประเมินผล ลัพธ์ที่ต่างกันไป ไม่จำเป็นต้องครบทั้ง 4 ประการก็ได้

นอกจากนั้น การทดสอบ การขึ้นระบบ และการนำ Model ไปใช้งาน จริงนั้น ระบบต้องเข้ากันได้กับ Data ขององค์กร โดยการขึ้นระบบแบ่งเป็น 3 ส่วน ได้แก่ ส่วนของการระบบพัฒนาระบบ ส่วนของการทดสอบระบบ และส่วนของการส่งมอบและสร้างความคุ้นเคยให้ผู้ใช้งาน จึงสะท้อนให้ผู้อ่านเกิดความเข้าใจในบริบทของการทำ Big Data ว่าเป็นรูปแบบการทำงานเพื่อรองรับความต้องการของธุรกิจโดยตรง

กล่าวโดยสรุปจากหนังสือ “Big Data Series II Think Like a Data Scientist คิดแบบนักวิทยาศาสตร์ข้อมูล” ลำดับแรกเริ่มจากมุมมองของผู้อ่าน พบว่า หัวใจสำคัญของการใช้ Big Data ไม่ใช่เพียงเครื่องมือ แต่เป็นกระบวนการคิด (mindset) ที่ต้องการใช้ Big data เข้ามาช่วยในการตัดสินใจและวางแผนธุรกิจหรือการตลาด เพราะตั้งแต่อดีตจนถึงปัจจุบัน องค์กรธุรกิจมักใช้ประสบการณ์ส่วนตัวในการตัดสินใจ เช่น แต่เดิมนักการตลาดจะคิดโปรโมชั่นต่าง ๆ ออกมาจากข้อมูลเป็นส่วนน้อย สำหรับปัจจุบันการทำการตลาดหรือแม้แต่โปรโมชั่นนั้น ๆ สามารถใช้ data เป็นแกนหลักในการตัดสินใจได้มากขึ้น ด้วยขณะนี้ได้เข้าสู่ยุคเปลี่ยนผ่านจาก Digital สู่ Data นั้นหมายความว่า การเก็บข้อมูลเป็นไปได้ง่ายขึ้น รวบรวมข้อมูลได้มากขึ้น และสามารถเอามาใช้งานจริงได้มากขึ้นเช่นกัน แสดงให้เห็นถึงเป้าประสงค์ของผู้เขียนว่า ผู้เขียนเน้นเนื้อหาด้าน การวิเคราะห์เชิงลึกทางด้าน Big Data ที่คนทั่วไปมักมีมุมมองว่าเป็นเรื่องใหม่ ในทางกลับกันการวิเคราะห์ข้อมูลมีมายาวนาน แต่กลับได้รับความสนใจและความนิยมเมื่อไม่นานมานี้ ภายหลังที่เทคโนโลยีสามารถประยุกต์ใช้และสนองตอบการทำงานที่เชื่อมโยงกันมากขึ้น ซึ่งชี้ชัดว่าการคิดเชิงลึกต้องอาศัยพื้นฐานความรู้มากขึ้น โดยเฉพาะการทำโครงการให้มีความแตกต่างแต่สร้างโอกาสให้องค์กรธุรกิจประสบความสำเร็จได้ต่อไป สำหรับในบทส่งท้ายเล่มของหนังสือผู้เขียนได้ทิ้งท้ายในลักษณะเชิญชวนให้ผู้อ่านติดตาม Big Data Series III : Big Data in Real Cases ซึ่งเป็นเล่มต่อของหนังสือเล่มนี้อีกด้วย และกล่าวได้ว่า หนังสือเล่มนี้เหมาะอย่างยิ่งสำหรับบุคคลที่มีความสนใจใน Big Data นักวิชาการและนักวิจัย บุคลากรในองค์กรภาครัฐและเอกชน รวมทั้งผู้ประกอบการธุรกิจ โดยเฉพาะอย่างยิ่งผู้นำและนักบริหารในองค์กรที่กำลังมองหาแนวทางการปรับตัวและปรับเปลี่ยนองค์กรให้ทันยุคสมัยอาจนำการทำ Big Data จากหนังสือเล่มนี้ไปปรับใช้ในการวางแผนการดำเนินงาน และการกำหนดกลยุทธ์ขององค์กรได้อีกทางหนึ่ง